DOCUMENT RESUME

ED 144 409 FL 008 915

AUTHOR Hisama, Kay K.
TITLE Patterns of Various ESOL Proficiency Test Scores by Native Language and Proficiency Levels. Occasional Papers on Linguistics, No. 1.
INSTITUTION Southern Illinois Univ., Carbondale. Dept. of Linguistics.
PUB DATE Apr 77
NOTE 11p.; Proceedings of the International Conference on Frontiers in Language Proficiency and Dominance Testing (1st, Southern Illinois University, Carbondale, Illinois, April 21-23, 1977)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Cloze Procedure; *English (Second Language); Evaluation Methods; Foreign Students; Higher Education; *Language Proficiency; Language Skills; Language Tests; Standardized Tests; Student Placement; *Test Bias; Testing; *Test Interpretation; Test Reliability; *Test Validity
IDENTIFIERS Comprehensive English Language Test (CELT); Reading for Understanding Test New Cloze Test

ABSTRACT

A profile method was used to analyze the patterns of four English proficiency tests (Comprehensive English Language Test for Speakers of English as a Second Language: Structure, CELT: Listening, Reading for Understanding Test, and The New Cloze Test) regarding two examinee characteristics: their language proficiency levels and native language. One hundred thirty-six non-native students at the Center for English as a Second Language (ESL) at Southern Illinois University at Carbondale served as subjects. Two types of scores were derived for this purpose: percent scores and standard scores. There were group differences regarding both examinee characteristics. The use of multiple measures may be defensible at the present time largely because available English proficiency tests show some biases in sampling a trait called English proficiency of ESL students. Their use may tend to cancel out or nullify such biases. There is, however, an element of luck here and a better solution would be to select deliberately a test or tests that minimize such biases. (Author/CFM)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED144409

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

PATTERNS OF VARIOUS ESOL PROFICIENCY TEST SCORES BY NATIVE LANGUAGE AND PROFICIENCY LEVELS

Kay K. Hisama

Southern Illinois University

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
James E. Redden, Editor
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

A profile method was used to analyze the patterns of various English proficiency tests regarding two examinee characteristics: their language proficiency levels and native language. Two types of scores were derived for this purpose: percent scores (P-scores) and standard scores(S-scores). There were some interesting group differences regarding both examinee characteristics. Administrators and teachers of ESOL would develop some insight as they review the findings. However, the same method should be applied to a different sample of ESOL learners in order to cross-validate such findings. In conclusion, the use of multiple measures may be defensible at the present time largely because available English proficiency tests show some "biases" in sampling a trait called English proficiency of ESOL students.

Today, testing is an essential part of instruction. The most popular use of standardized tests is to group or place students for instructional purposes. Paper-and-pencil tests are widely used when there are a large number of examinees. There is no question that these tests greatly simplify testing processes while facilitating instruction.

The most important but difficult aspect of testing is its interpretation. The task is a formidable one indeed, because test scores are not "pure" measures of a certain ability which the test intends to measure. A host of irrelevant factors tend to entangle with any test score. The general purpose of this study is to shed some light on the aspect of English language proficiency tests given to a large number of students of English as a second language.

There are three major English proficiency tests for non-native speakers on the market. All of these tests consist of several subtests. In the case of the Test of English as a Foreign Language (TOEFL), there are five subtests: Listening Comprehension, English Structure, Vocabulary, Reading Comprehension, and Writing. The term "battery" is conventionally applied to such a set of separate tests to be administered to a group of individuals. It should be recognized, however, that even though the identity of each measure is preserved up to the end, all related measures are ultimately combined in the decision to take(or recommend) action on the basis of the available data. This practice of multiple measurement is believed to be a sound one in the light of much

FL008915

previous research. For example, the results of several tests are likely to give a better prognosis of college achievement than any one test alone. The most obvious reason for the superiority of multiple measurement is that the complexity of any human ability requires a wide range of item samples. Many individual tests tend to fail to measure any complex human ability.

This study attempts to take a closer look at three standardized English proficiency tests which are currently being used at the Center for English as a Second Language (CESL) at Southern Illinois University at Carbondale for the purpose of grouping entering students. In addition, one experimental test was also administered to facilitate the interpretation of these test scores. Questions were raised as to in what fashion, if any, levels of an examinee's language proficiency and his characteristics related to his native language, would interact with these tests.

METHOD

Subjects

A total of 136 non-native students at the Center for English as a Second Language (CESL) at Southern Illinois University at Carbondale served as subjects. Out of eight sessions in a year, the following sessions were sampled: October, November, and January. Two criteria for the subjects were that the subject had to be a new entering student to the center, and that he or she had not been in the United States longer than one month at the time of testing. The test scores of those who met the above criteria were retained for analyses. They represented 12 different languages, Arabic (N=29), Farsi-Persian (N=49), Spanish (N=27), French and one of the African languages (N=10), Japanese (N=7), and the remaining represented by less than seven subjects each.

Test Instruments

Four tests were administered to the subjects upon their entrance to the center: (1) A Comprehensive English Language Test for Speakers of English as a Second Language (CELT): Structure, (2) CELT: Listening, (3) Reading for Understanding Test (RFU), and (4) The New Cloze Test(NCT). The first three tests are of multiple choice format where the examinees are required to choose one of four alternatives. However, the three tests differ from each other in several respects.

First, two CELT tests were specifically developed and validated with non-native speakers of English while RFU was constructed as a placement test for a remedial reading program. The program is designed for those who are in the intermediate elementary school levels through the high school and college levels.

Second, the dominant mode of language used for the two CELT tests is conversational. The test contains a number of colloquial expressions of which a large number of foreign students have never learned. On the other hand, the register of the RFU is from written materials which are likely to be used at school. The bulk of questions in the RFU require that the subject have a good conceptual understanding of various school subjects such as general science, language arts or social studies.

Third, the total number of items for each of the three tests differ considerably, ranging from 50 for the CELT-L to 100 for the RFU. Accordingly, the timing of the three tests varied. The CELT-L requires all examinees to proceed at the same speed while the others have a fixed time limit for the entire test. These differences in the manner of timing, in the total number of test items, together with the difficulty level of each test tend to increase the effects of guessing on test scores somewhat differently.

The fourth test, the New Cloze Test, radically differs from the above three tests in that there is no random guessing involved in the test. Although the NCT uses a fill-in-the-blank format, it differs considerably from the variety of tests known as cloze tests. The test was designed to measure an aspect of ESOL proficiency that is important in meaningful human communication for a group of examinees of various proficiency levels.

Method of Analysis

The method used in this study is a "profile" method. A "profile" is a graphic summary of the results of multiple measurement. The principle advantage of the profile method is its apparent simplicity in presenting the results of testing at the nontechnical level. One major drawback of the method stems from unreliability of individual test scores. However, all tests used for this study demonstrated extremely high reliability and predictive validity. (For these and other statistical data, see Hisama,1976.) Therefore, the resulting profiles of these test scores can be interpreted with a reasonable degree of confidence.

Since the four tests have a different number of items, raw scores (the number of correct answers) were transformed to two types of "comparable scores." One is the percent score or P-score which indicates the percent of correct answers. The use of P-scores is the simplest way to examine the relative difficulty of various tests for the entire group or for each subgroup. However, P-scores of different tests have different means and variations within the test. This would limit the generalization of findings regarding the profiles of these tests by various subgroups. Therefore the standard scores were also derived from the raw scores. The standard scores (S-scores) used for the analyses have standard values of mean of 50, and a standard deviation of 10. Both types of scores were used in order to construct profiles regarding two characteristics of the subjects: their language proficiency levels and their native language.

Four figures were drawn by plotting both the P- and S- scores on the four tests obtained by different subgroups based on two examinee characteristics. Composite lines, which show the weighted average of three multiple choice test scores were also drawn for each subgroup to facilitate the interpretation of multiple measurement.

RESULTS AND DISCUSSION

Profiles of Various Test Scores by EFL Proficiency Levels

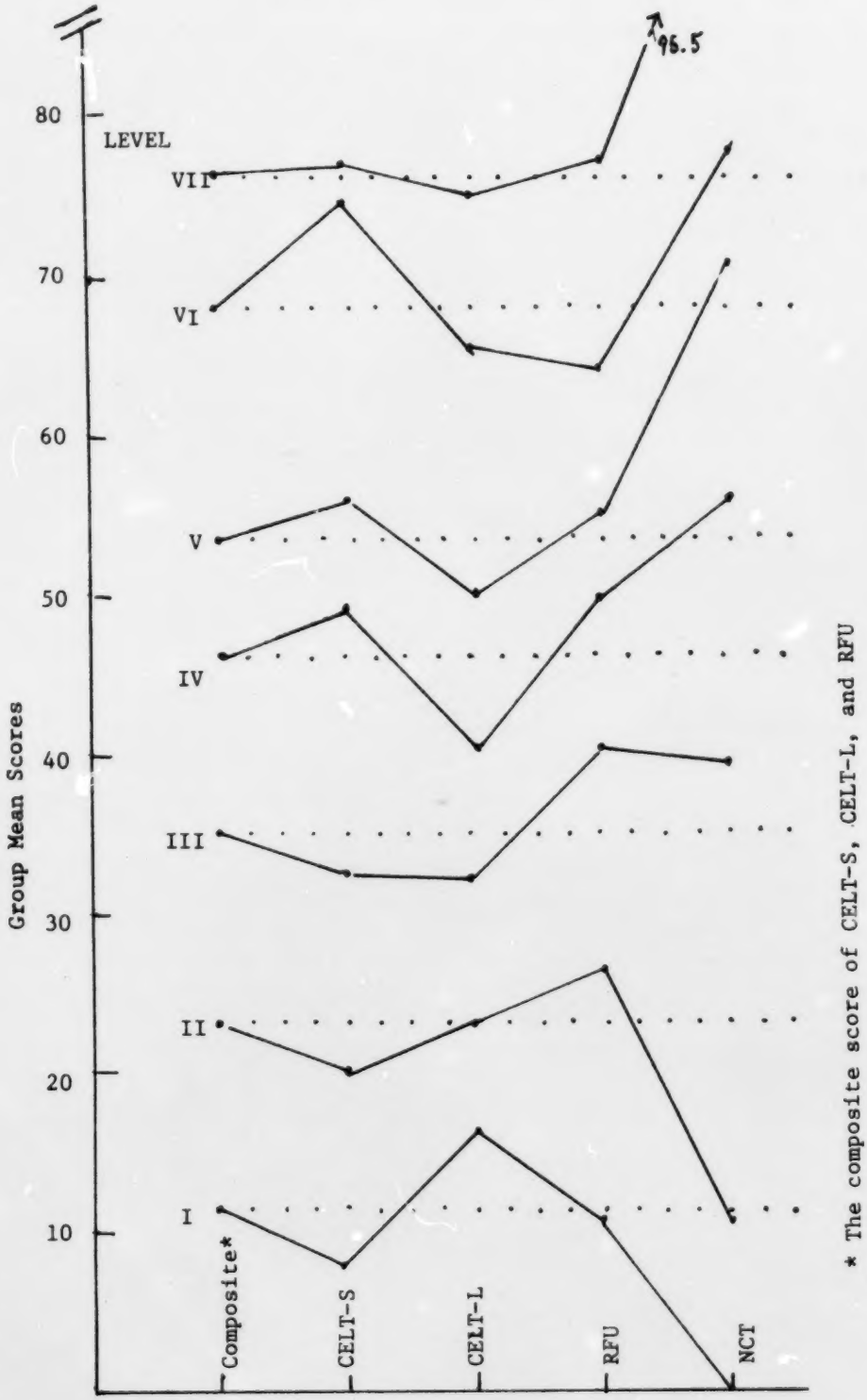

FIGURE 1 PROFILES OF DIFFERENT LANGUAGE PROFICIENCY LEVELS BASED ON PERCENT SCORES

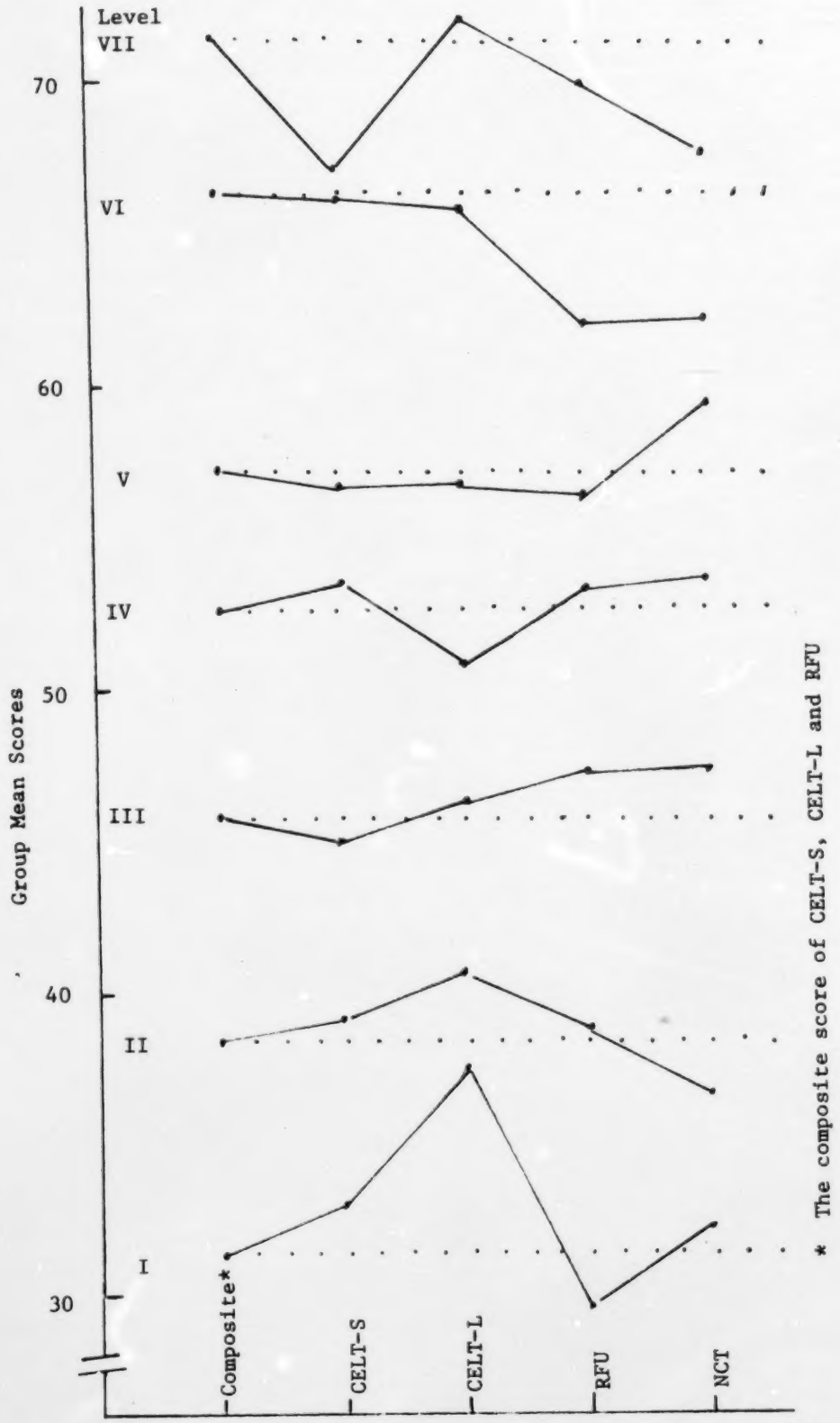

FIGURE 2 PROFILES OF DIFFERENT LANGUAGE PROFICIENCY LEVELS BASED ON STANDARD SCORES

Figure 1 shows the mean percent scores obtained by seven prociciency groups.[1] Moderate to large score differences in all test scores among the seven groups indicate the high predictive validity of these tests for placement purposes. Turning to the score patterns within each group, the deviation of four test scores from the composite score line indicates that there is some variation among the four tests in their demonstrated difficulty for each group. The relative score patterns for each test are summarized below.

CELT-Structure: The mean P-scores on CELT-S by the seven groups show that while three lower groups obtained somewhat lower means than their respective composites, the trend was opposite in the cases of the next three higher groups. CELT-S fails to discriminate between the two highest levels. This might indicate that the test fails to measure an aspect of language proficiency possessed by the highly proficient students. Figure 2 confirms this observation.

CELT-Listening: While the mean P-score on CELT-L for the entire group is the lowest of the three multiple choice tests, there are some differences in the relative difficulty of CELT-L for each group. The mean P-scores on CELT-L by levels 3 through 6 are somewhat lower than their mean scores on other tests. This contrasts with those by two low level groups and that by level 7. These score patterns on CELT-L might be because of the different manners of timing these tests. Since the test forces the examinees to proceed at the same speed, all examinees are given a chance to examine all items and mark them even though they are not sure of the correct answers. It follows, then, that those of lower ability groups have more items in which they can guess randomly and increase the probability of getting extra points. The mean score on CELT-L by level 7, the highest level, is impressive considering that they had the least number of items in which they could guess randomly and could receive extra points. Figure 2 confirms this observation.

Reading for Understanding Test: RFU discriminates among all proficiency groups including the two highest groups.

The New Cloze Test (NCT): Unlike the above three tests, the NCT does not have the effect of random guessing. The adverse effects of random guessing is the most serious when the examinees have widely differing backgrounds, and when there are a large number of test items which are too difficult for some of them. The overall pattern shows that the NCT contains few items which are inappropriate for testing this group, indicating that the test is not only highly valid but also highly efficient. It is interesting to note that scores on NCT are relatively lower than those on other tests for the two lowest groups. The trend is opposite for the other groups. This might be accounted for by the different effects of guessing by different proficiency groups.

---

1. The six proficiency levels at CESL vary from Level 1 -- knows a little English, Level 5 -- can converse fluently on most subjects to Level 6 -- a course designed for those who want to study at graduate schools; Level 7 -- above Level 6.

Discussion

Based on the above observations, there seem to be several major components of ESOL proficiency. The most basic of such components is the understanding of the basic structure of the language as it is represented by the CELT-S test. However, as a learner improves in his proficiency, his conceptual understanding of the content of some meaningful materials becomes important. This was demonstrated by the higher level students in their continual growth as they were reflected by their test scores on the NCT. The results of RFU also seem to confirm this argument because many questions of RFU require conceptual understanding of reading materials commonly used at schools.

The results on CELT-Listening may indicate that the speed of understanding a given message is another important component of learning language. The measurement of this component is important to identify highly proficient students. However, careful consideration should be given in order to control the sources of error which are inherent to this type of test.

Profiles of Various Test Scores by Native Language

Figure 3 shows the profiles of mean percent scores obtanined by five different language groups.[2] Figure 4 presents those of standard scores for the same groups.

The overall pattern shows that there are small to moderate differences among the group means on the four tests as well as composite. A host of factors may account for such differences: availability of college education in their native countries, availability of scholarships or private funds, geographical distance, international relationships with the USA, or a combination of them. For example, when their government has a limited amount of funds, it can give scholarships only to elite students. On the other hand, if the government has unlimited funds and there are not enough colleges, it would send even those who are average students. There is also sample variation of subjects. Therefore, the most interesting research question regarding various language groups is to examine the extent to which their test scores differ from each other within each language group.

Both figures 3 and 4 show that in the CELT-S, little difference exists among the three highest groups--Asian, African, and Persian, even though there are considerable differences in the other tests. Considering that scores on the other tests by the Persian group are much lower than the two other groups, it is reasonable to say that the CELT-S "advantageous" to the Persian group.

The relatively low score on the CELT-L by the African group indicates that this test is to their disadvantage. It was confirmed that they have the opportunity to listen to British broadcasting in their country. Their unfamiliarity with American English made the CELT-L difficult for this group. It should also be noted that the mean score on the CELT-S by the African group is also relatively lower even though it is not as low as that on the CELT-L. Considering extremely high scores on

---

2. The Asian group(N=18) includes the speakers of four languages: Japanese, Chinese, Vietnamese and Thai. Other groups were: Arabic(N=29), Farsi-Persian(N=49), Spanish(N=27), and African(N=10).

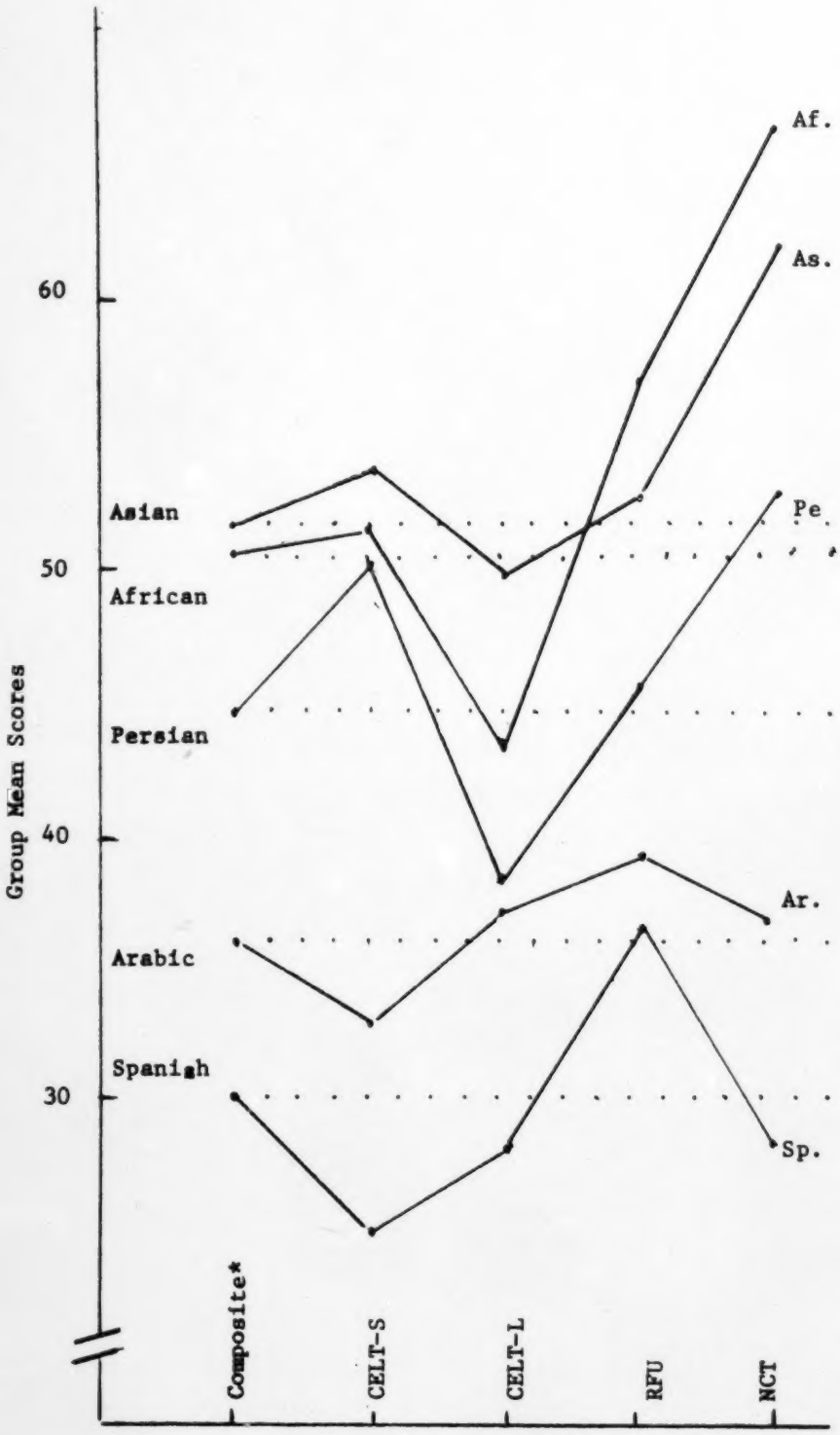

FIGURE 3 PROFILES OF DIFFERENT LANGUAGE GROUPS BASED ON PERCENT SCORES

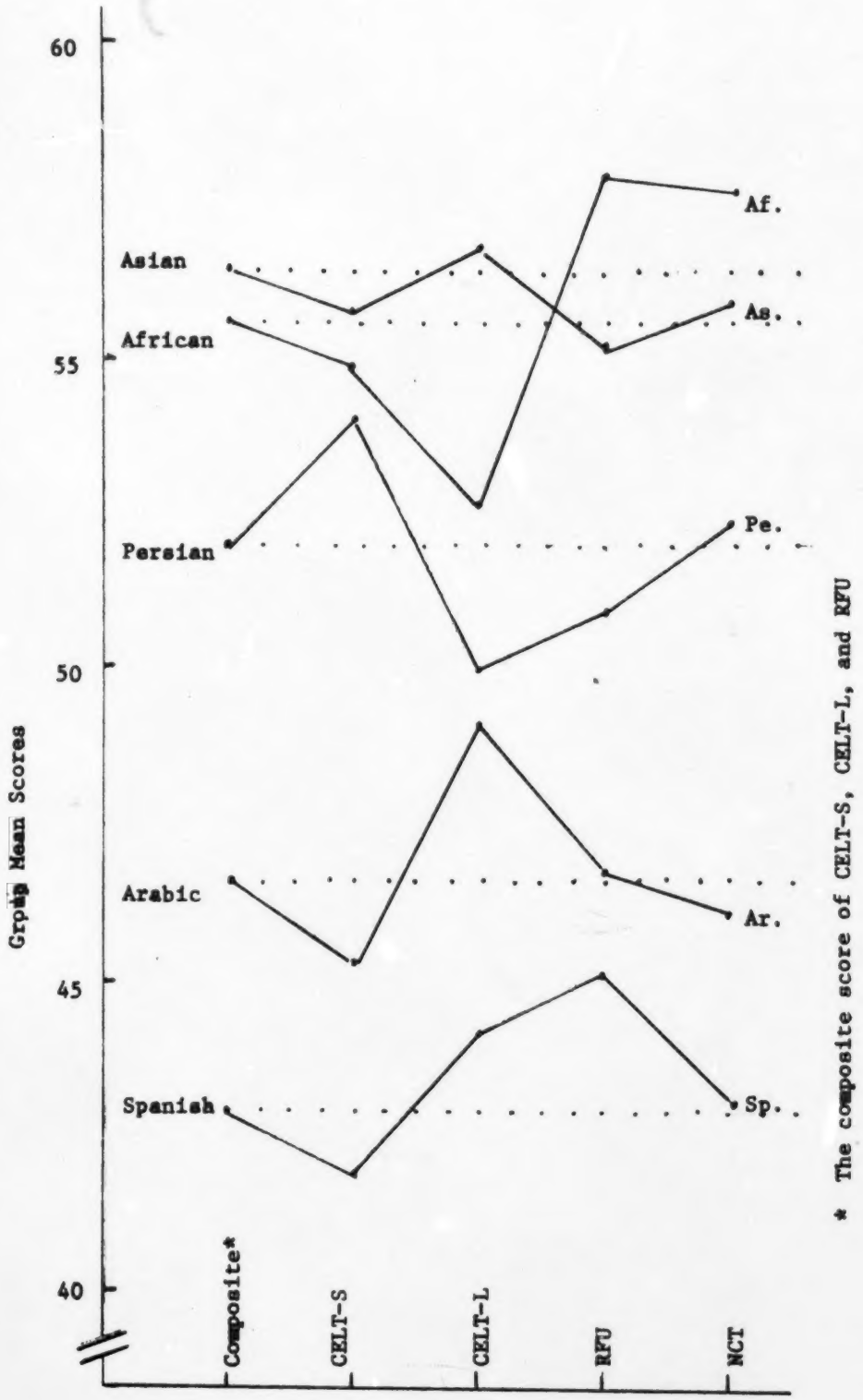

FIGURE 4 PROFILES OF DIFFERENT LANGUAGE GROUPS BASED ON STANDARD SCORES

the RFU and the NCT by this group, the results on the two CELT tests seem to show that both CELT tests contain a large number of colloquial expressions of American English with which the African group is not familiar.

Test scores obtained by the Asian group show that their performance on the different tests is well balanced. Their performance on the listening test is impressive compared to those by the two other higher groups on the same test. This may indicate that many Asian countries have been under the influence of American culture including its language. They have opportunities to listen to American English through broadcasting or records and tapes, recorded in American English.

A relatively higher score on the RFU by the Spanish group is interesting. However, it is somewhat difficult to interpret this result because the RFU contains a large number of difficult items, which, in turn, magnifies the undesirable effect of random guessing on this test in the case of a low ability group.

SUMMARY AND CONCLUSIONS

A profile method was applied to analyze the results of the four tests. It was revealed that there are somewhat different score patterns depending on two examinee characteristics: their proficiency levels and their native language. In some cases a subgroup scored relatively better on one test but not on another as their scores on each test were compared with the composite. Score pattern differences among the different subgroups may be largely the results of the "biases" in sampling differences in the tests such as the number of items, the effect of random guessing, the register of test items, accent of the speakers on the listening test, and so on. Thus, multiple measures of ESL proficeincy may be defended by arguing that their use would tend to cancel out or nullify such biases. However, there is an element of luck here, and a better solution would be to select deliberately a test or tests that minimize such biases.